Sporting Goods Late Delivery Risk

Tyler Mann Predictive Modeling 202140 BANA - 7365 – 040

**QUESTION/TOPIC:**

As someone who works in supply chain, I am always curious about late deliveries, why they happen, and what the cause of late deliveries may be. With interest in sports, health, and apparel continuing to grow around the world, supply chain efficiency is more important than ever before. From the standpoint of retail corporations (Walmart, Target, etc.…) and 3PL’s (warehousing/transportation), it is important that they all work together to provide on-time service to their customers. Since the emergence of Amazon, on-time delivery has become more important than ever before to compete with “Prime” delivery and its efficiency. In my own company I have seen how late delivery risk is more of a hot topic than it ever was when I first joined the company. Retailers set on-time delivery metrics (upwards of 95%) that we must meet each week or face heavy fines. It is a way for them to make more money from their suppliers and force supply chain to the forefront of importance to managers along with sales and marketing. The importance of an effective and efficient supply chain is pivotal to a businesses’ success, so I decided to analyze sporting goods late delivery risk on a global scale for some of the most popular sporting good’s items. The focus will be on late deliveries and trying to find correlation for why there’s so many late deliveries as you will see below based on the data.

**LITERATURE REVIEW:**

“This dataset was used by the company DataCo Global, which allows the use of machine learning algorithms and R/Python software. Areas of important registered activities include provisioning, production, sales, and commercial distribution. It also allows the correlation of structured data with unstructured data for knowledge generation.” - Fabian Constante, Fernando Silva, António Pereira.

The peer reviewed article named “Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities,” which discusses big data analytics in supply chain management and how it is receiving more attention. Therefore, to increase the precision of demand forecast, supply chain data shall be carefully analyzed to enhance knowledge about market trends, customer behavior, suppliers, and technologies. Extracting trends and patterns from such data and using them to improve accuracy of future predictions can help minimize supply chain costs Analysis of supply chain data has become a complex task due to (1) increasing multiplicity of SC entities, (2) growing diversity of SC configurations depending on the homogeneity or heterogeneity of products, (3) interdependencies among these entities (4) uncertainties in dynamical behavior of these components, (5) lack of information as relate to SC entities, (6) networked manufacturing/production entities due to their increasing coordination and cooperation to achieve a high level customization and adaptation to varying customers’ needs, (7) and finally the increasing adoption of supply chain digitization practices (and use of Blockchain technologies) to track the activities across supply chains. This combination of data sources used in SC demand forecasts, with their diverse temporal and spatial attributes, places a greater emphasis on use of big data analytics in supply chains, in general, and demand forecasting efforts. – Mahya Seyedan & Fereshteh Mafakheri

**DATA:**

Appendix 1:

|  |  |
| --- | --- |
| Fields | Description |
| Type | : Type of transaction made |
| Days for shipping (real) | : Actual shipping days of the purchased product |
| Days for shipment (scheduled) | : Days of scheduled delivery of the purchased product |
| Delivery Status | : Delivery status of orders: Advance shipping, Late delivery, Shipping canceled, Shipping on time |
| Late\_delivery\_risk | : Categorical variable that indicates if sending is late (1), it is not late (0). |
| Market | : Market to where the order is delivered: Africa, Europe, LATAM, Pacific Asia, USCA |
| Sales | : Value in sales |
| Order Status | : Order Status: COMPLETE, PENDING, CLOSED, PENDING\_PAYMENT, CANCELED, PROCESSING, SUSPECTED\_FRAUD, ON\_HOLD, PAYMENT\_REVIEW |
| Shipping Mode | : The following shipping modes are presented: Standard Class, First Class, Second Class, Same Day |

Source of the dataset: [Mendeley Data - DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS](https://data.mendeley.com/datasets/8gx2fvg2k6/5)

For data cleaning I uploaded the dataset, read it, and then ran a df.head() to get a visual of the data columns. I then ran a df.columns to see my columns in their entirety and ease of selecting the columns. According to df.shape I have 53 columns and 180,519 rows of data in total. I then ran a df.isnull() to get the sum of all null values. Those that had null values would later be deleted from the columns. I choose specific columns to look at different graphs and analytical numbers.

Lastly, I created dummy variables for “Type”, “Delivery Status”, “Market”, “Order Region”, “Order Status”, and “Shipping Mode”, eventually dropping “Order Region” because it had too many variables.

Appendix A:

Here is where I notice the large number of late deliveries averaging out to just over 54%, which is an erratic percentage. I then knew I wanted to continue to look at late deliveries and run a model based on variables that are impacted by it. I drop columns that are not needed and continue to analyze the data, looking at type and info on the different variables. Lastly, I run a df.describe() to get the descriptive statistics for 29 different columns. The descriptive statistics gave me the mean days for real vs scheduled shipments of 3.50 vs 2.93 which doesn’t seem bad on average but would depend on the customer’s target metrics. Again, you can see that late delivery risk was high at 54% meaning that if it was at risk of being late, it was essentially late. For late delivery risk, a standard deviation of 0.50 means that about half are close to the mean.

**MODEL:**

I have chosen to use a logistic regression model which relies on concepts such as conditional probability and odds ratio. Logistic regression is used when the output variable is a binary or categorical variable rather than a continuous variable for linear regression. Since we are looking at late delivery risk (0 or 1), no or yes, according to “Learning Predictive Analytics with Python.” – Ashish Kumar

In “Predicting On-time Delivery in the Trucking Industry”, on-time delivery is a key metric in the trucking segment of the transportation industry. If on-time delivery can be predicted, more effective resource allocation can be achieved. Their research focuses on building a predictive analytics model, specifically logistic regression, given a historical dataset. – Rafael Durarte & Kenneth Ohlund

Also, the importance of logistic regression is stated in the article “Logistic regression in modeling and assessment of transport services”, states: “There are different methods for such an analysis. This article proposes logistic regression. The research was conducted based on a distribution and trade company dealing with the supply of automotive spare parts. As the most profitable group of customers are local car repair shops, it was this group that was subject to analysis. The quality-of-service assessment was considered from the point of view of delivery time. The dichotomous form of the predictor taking two values - late and on-time delivery - was determined. From among the possible ones, regressors whose influence was statistically significant and whose modification was possible were selected. The research showed which of them (and how strongly) affect the dependent variable, which allowed for modification of strategy and implementation of new solutions increasing the number of satisfied customers.” – Anna Borucka

My dependent variable is “Late\_delivery\_risk” while I have various independent variables that I use to witness the results: 'Days for shipping (real)', 'Days for shipment (scheduled)', ‘Delivery Status’, ‘Market’, ‘Order Status’.

**ANALYSIS:**

Appendix B – B3:

This is just a look at different variables in the beginning to see the metrics for my own understanding of the data.

Appendix C:

Here is where I noticed the high percentage of late deliveries vs shipping on time. I then wanted to see if anything correlated with late deliveries.

Appendix D:

I look at the type and information for the dataset. I then print a df.describe() to see the statistics for the numerical variables. Here you can see the mean for days shipping real vs scheduled, sales averages, the high percentage of “Late\_delivery\_risk”, average price and profit, etc. The focus for me becomes the late delivery risk and how often those risks become reality over 50% of the time.

I want to see if there’s correlation between any of the data that I could work with, so I choose some columns to analyze, and it looks that “Days for shipping (real) correlate the most out of the selected columns. It makes sense as that column has to do with actual days till it shipped. I have a few views that show different correlations.

Appendix E:

A contingency table is a representation of the frequency of observations falling under various categories of two or more variables. It comes in a matrix form and essentially contains the frequency of occurrences for the combination of categories of two or more variables. In the appendix, we are looking at “Late\_delivery\_risk” vs real and scheduled shipments. Zero means no late delivery risk and one means there is a late delivery risk. For “Days for shipping (real), we can see that risk is definitely a problem the more days shipments take. Same thing for “Days for shipment (scheduled), way too many orders fall under late delivery risk. There is also bar charts showing the level of risk for both real and scheduled shipments.

Appendix F:

Now we start our process for logistic regression by first picking our independent variables to go along with our dependent variable of “Late\_delivery\_risk”. I then start the test data and pick our categorical variables that will receive dummy variables. I notice that we have too many variables so now I pick out more than need to be deleted (dropped) from the test.

Appendix G:

I then add the columns that will be tested in the logistic regression model and run those separately based on category. Those categories are “Order Status”, “Market”, “Delivery Status”, and “Days for shipping/shipment (real vs scheduled). We will look at each of them with “Late\_delivery\_risk” as our independent variable.

Appendix H:

The first is looking at the dependent variable “Late\_delivery\_risk” and the various independent “Order Status” variables. I see that the p-values of CLOSED, COMPLETE, ON\_HOLD, PAYMENT\_REVIEW, PENDING, PENDING\_PAYMENT, and PROCESSING are significant because they are lower than 0.05. For example, when comparing two order statuses that are PENDING who differ by one unit associated with risk, the order with risk PENDING order will, on average, have 0.32 more chance for late delivery risk. This means there is strong evidence of a logistic association between the variables “Late\_delivery\_risk” and “PENDING”. A Pseudo R-squared of 0.05137 means that 5% of the variation in the variable “Late\_delivery\_risk” is explained by the variable “Order Status\_PENDING” and so on for each variable. The standard error is high for CANCELED and SUSPECTED\_FRAUD meaning they are an inaccurate representation of the true population mean and are not significant. My ROC curve has a false positive rate of 90.79%, which means that often an actual negative instance will be classified as positive. Meaning we almost have no predictive power (random guessing).

Appendix I:

The last one looking at “Late\_delivery\_risk” as your dependent variable and “Days for shipping (real)” and “Days for shipment (scheduled)” I found to be interesting. Their p-values are less than 0.05 so they are significant. Meaning when comparing two shipments of (real) one order on average, will have a 3.63 more chance for late delivery risk while (scheduled) has a -4.23 chance for late delivery risk. The ROC curve has a much higher accuracy of 0.977 and a false positive rate of 5.13%. This means that we have a high true positive, so this is a good model.

**CONCLUSION:**

My findings are that “Late\_delivery\_risk” is a big issue that needs to be resolved between corporate, warehousing, and transportation. The payment process is significant and those orders that are not complete could be causing orders to deliver late or not at all. The number of days for real vs scheduled are also significant and play a huge role whether an order will deliver late or not. Management will need to look at these findings to figure out how to decrease those late deliveries and increase on-time shipments. Half of them being late will not help retailers compete with Amazon so a team needs to be put together to go over the data and find ways to resolve the payment issues.

Work Cited:

Seyedan, M., Mafakheri, F. Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. *J Big Data* **7,**53 (2020). <https://doi.org/10.1186/s40537-020-00329-2>

[Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities | Journal of Big Data | Full Text (springeropen.com)](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00329-2#Bib1)

Dziubak, Tadeusz & Wysocki, Tadeusz & Dziubak, Sebastian. (2021). Selection of vehicles for fleet of transport company on the basis of observation of their operational reliability. Eksploatacja i Niezawodnosc - Maintenance and Reliability. 23. 184-194. 10.17531/ein.2021.1.19.

[(PDF) Logistic regression in modeling and assessment of transport services (researchgate.net)](https://www.researchgate.net/publication/339328045_Logistic_regression_in_modeling_and_assessment_of_transport_services)

[PlumX - DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS](https://plu.mx/plum/a/?mendeley_data_id=8gx2fvg2k6&theme=plum-bigben-theme)

Alcoba, R. D., & Ohlund, K. W. (n.d.). *Predicting On-Time Delivery in the Trucking Industry*. Transportation, Predictive Analytics. https://ctl.mit.edu/sites/ctl.mit.edu/files/theses/Duarte%2BOhlund%20executive\_summary.pdf.

Constante, Fabian; Silva, Fernando; Pereira, António (2019), “DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS”, Mendeley Data, V5, doi: 10.17632/8gx2fvg2k6.5

[DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS | Kaggle](https://www.kaggle.com/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis)

[E-Commerce Multi- Output Models {Project\_CSE07} | Kaggle](https://www.kaggle.com/sukanthen/e-commerce-multi-output-models-project-cse07)

[Comparison of ML models with RNN | Kaggle](https://www.kaggle.com/jaswanthhbadvelu/comparison-of-ml-models-with-rnn)

[Dataco Analyse | Kaggle](https://www.kaggle.com/mowniesh/dataco-analyse)